

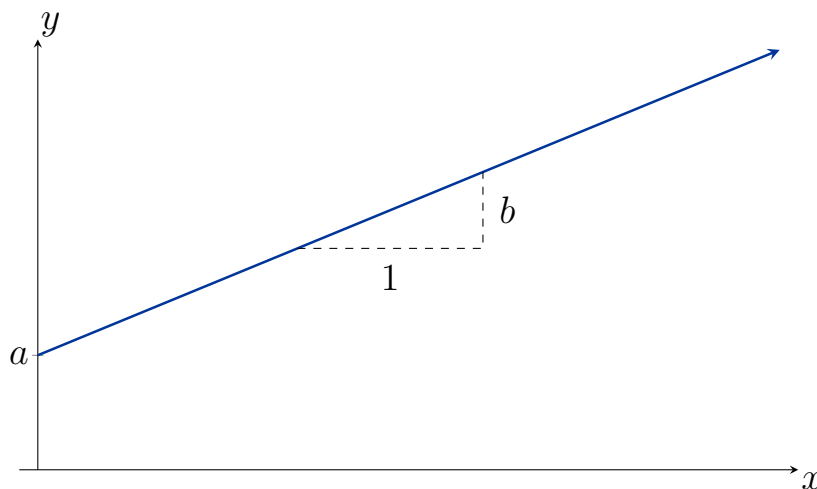
4.3: Modeling Linear Trends

Definition.

The **regression line** is a model used for making predictions about *future* observed values. The equation of the regression line is

$$y = a + bx$$

where a is the **y -intercept** (when $x=0$) and b is the **slope**.



The input variable x is known as the

- Independent variable
- Predictor variable
- Explanatory variable

The output variable y is known as the

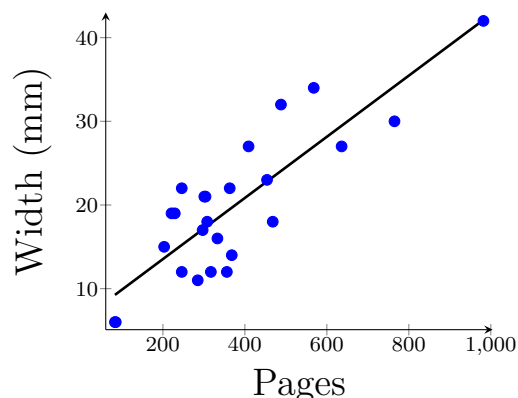
- Dependent variable
- Predicted variable
- Response variable

Example. Below is a scatterplot comparing number of pages a book has against the width of the book. Interpret the intercept and the slope of the regression line.

Intercept: Book with no pages (cover only) is 6.22mm wide

Slope: When we compare to a book with one more page, that book will be 0.0366mm wider

$$\text{Predicted Width} = 6.22 + 0.0366 \text{ Pages}$$

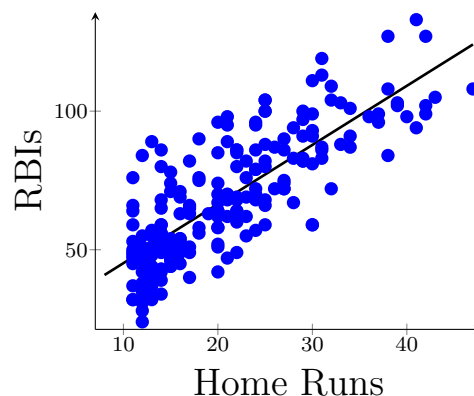


Example. Below is a scatterplot comparing the number of home runs and RBIs in the 2016 season. Interpret the intercept and slope of the regression line.

Intercept: 23.84 RBI when number of HR=0

Slope: When comparing one player to another with one additional HR, the other player's RBI is expected to be 2.13 higher

$$\text{Predicted RBI} = 23.84 + 2.13 \text{ HR}$$



Definition.

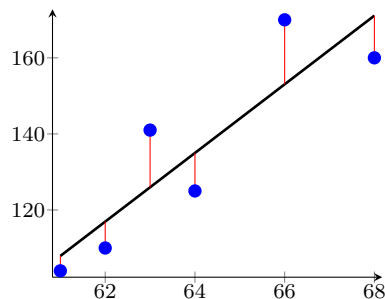
Now we define the formula of the regression line:

$$y = a + bx$$

Where

$$b = r \frac{s_y}{s_x} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

These formulae minimize the residual error: [Try this!](#)



Example. Below are the heights and weights of six women:

Heights	61	62	63	64	66	68
Weights	104	110	141	125	170	160

From this we get

$$\bar{x} = 64$$

$$s_x = 2.608$$

$$\bar{y} = 135$$

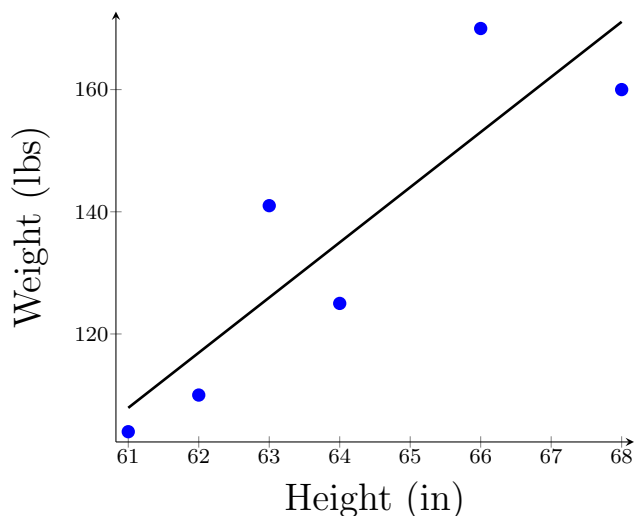
$$s_y = 26.728$$

$$r = 0.881$$

Find the equation of the regression line.

$$b = 0.881 \cdot \left(\frac{26.728}{2.608} \right) = 9.03$$

$$a = 135 - 9.03(64) = -442.92$$



Example. Open the `popdensity_and_crime` dataset in StatCrunch, use the “Simple Linear” tool under the `Stat>Regression` menu to find the regression line for the following columns. Interpret the slope and intercept where appropriate.

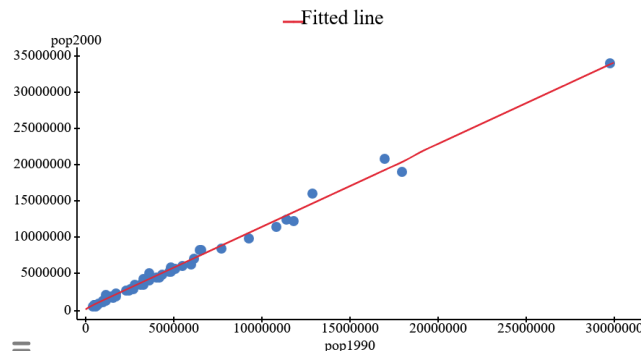
the `pop1990` and `pop2000` columns,

$$\text{pop2000} = 12266.759 + 1.1295246 \text{ pop1990}$$

 $R(\text{correlation coefficient}) = 0.99649554$

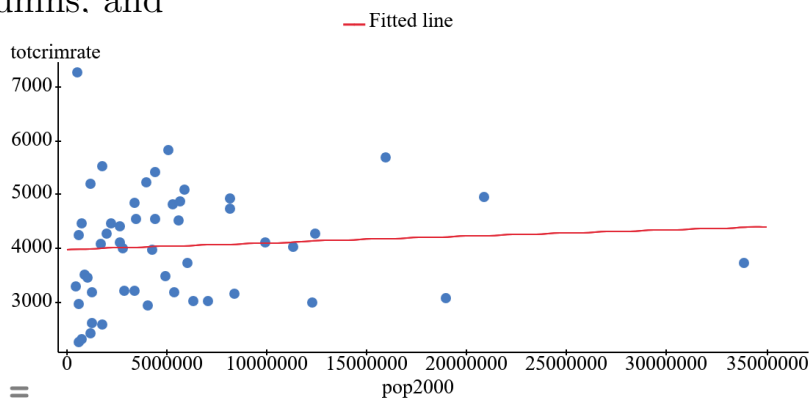
Intercept: If `pop=0` in 1990, it's 12,267 in 2000
 (Not appropriate since data is 454k-30mil)

Slope: When comparing pop in 1990 where one group has 1 more person, that group is expected to have 1.13 more people in 2000



the `pop2000` and `totcrimrate` columns, and

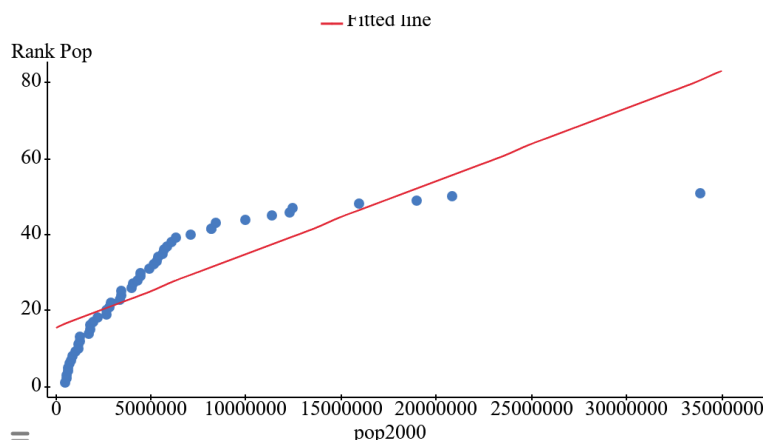
Low correlation coefficient (~0.07)
 Looks nonlinear
 Linear regression is not appropriate here



the `pop2000` and `Rank Pop` columns.

Large correlation coefficient (~0.8),
 but a nonlinear trend

Linear regression is not appropriate



4.4: Evaluating the Linear Model

Guidelines:

- Don't fit linear models to nonlinear associations!
- Correlation is not causation
- Beware of outliers (a.k.a. **influential points**)
- Don't extrapolate (make predictions beyond the range of the data)

Definition.

The **coefficient of determination** is the correlation coefficient squared:

$$r^2$$

This is sometimes also called ***r*-squared**.